PCA of Σ: performing the computations we obtain:

$$\begin{cases} \lambda_1 = 100.16 \;, & \underline{e}_1 = (0.04,\ 0.999)'. \\ \lambda_2 = 0.84 \;, & \underline{e}_2 = (0.999,\ -0.04)'. \end{cases}$$

Then we have:

$$\begin{cases} Y_1 = 0.04 X_1 + 0.999 X_2. & \to \text{basically } X_2. \\ Y_2 = 0.999 X_1 - 0.04 X_2. & \to \text{basically } X_1. \end{cases}$$

An example is the dataset of the competitions in athletics of various nations. Using PCA on that dataset, we would find that the PC_1 will be the marathon.

PCA of S:

$$\begin{cases} \lambda_1 = 1.4 \;, & \underline{e}_1 = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)' \\ \lambda_2 = 0.6 \;, & \underline{e}_2 = \left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right)'. \end{cases}$$

$$\begin{cases} Y_1 = \frac{1}{\sqrt{2}} Z_1 + \frac{1}{\sqrt{2}} Z_2 = \frac{1}{\sqrt{2}} \left(\frac{X_1 - \mu_1}{\sqrt{\sigma_{11}}} + \frac{X_2 - \mu_2}{\sqrt{\sigma_{22}}} \right) \\ Y_2 = \frac{1}{\sqrt{2}} Z_1 - \frac{1}{\sqrt{2}} Z_2 = \frac{1}{\sqrt{2}} \left(\frac{X_1 - \mu_1}{\sqrt{\sigma_{11}}} - \frac{X_2 - \mu_2}{\sqrt{\sigma_{22}}} \right). \end{cases}$$

The linear combinations forming Y_1, Y_2 are very different from the ones of before!

THM: if data is very influenced by units of measure etc., then use the standardized data $\underline{Z}$ and apply PCA on it.

Summarizing Sample Variation by Principal Components:

Usually, $\underline{\mu}$ and Σ are unknown. But we have data!

$$\mathbb{X} = \begin{bmatrix} \underline{x}_1' \\ \vdots \\ \underline{x}_n' \end{bmatrix} \text{ data}, \qquad \underline{x}_i \text{ realization of } \underline{X}_i, \quad \underline{X}_1, \dots, \underline{X}_n \text{ iid} \sim \underline{X}.$$

Hence we will use data to estimate $\underline{\mu}$ and Σ.
In particular, Σ is estimated by S and μ by $\bar{x}$.

$\Rightarrow$ we perform PCA on S:

$$S = \sum_{i=1}^{p} \lambda_i \underline{e}_i \underline{e}_i' \quad \Rightarrow \quad PC\ Y_i : \text{projection on } \underline{e}_i.$$

$$\underline{x}_i \xrightarrow{PCA} \underline{y}_i = \begin{pmatrix} \underline{e}_1' \underline{x}_i \\ \vdots \\ \underline{e}_p' \underline{x}_i \end{pmatrix}$$

scores of $\underline{x}_i$ (i.e. projection of $\underline{x}_i$ in the space of eigenvectors P)

Hence:

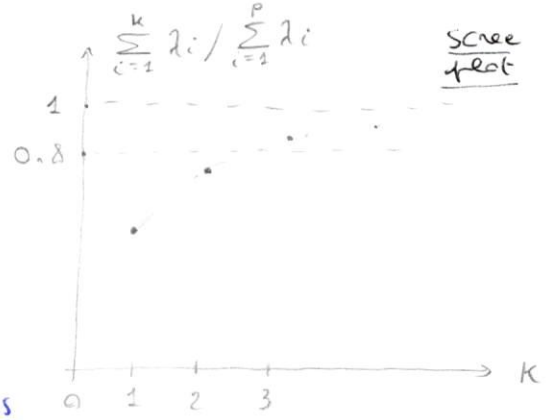
$$\mathbb{X} = \begin{bmatrix} \underline{x}_1 \\ \vdots \\ \underline{x}_n \end{bmatrix} \xrightarrow{PCA} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} Y_1 & Y_2 & \cdots & Y_p \\ y_{11} & y_{12} & \cdots & y_{1p} \\ \vdots & & & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{bmatrix}.$$

- Number of principal components:

Suppose that you set a threshold:

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \geq \text{threshold} \quad (\text{e.g. } 0.8).$$

In this way, we can reduce the analysis to the first k principal components.



Observation 1: PCA for categorical variables is called correspondance analysis.
It is performed on the table of joint frequencies (contingency tables).

Observation 2: look also at the smallest λ_i.
If $\lambda_p \simeq 0$, it means that there is a linear relationship between $X_1, \dots, X_k$.

We now look at a more "mathematical fashion" for the derivation of PCA.

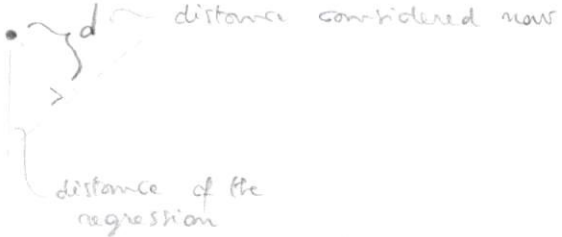
Derivation of PCA: Optimal Orthonormal Basis:

$\underline{x}_1, \dots, \underline{x}_n \in \mathbb{R}^p$.

Problem: find a linear space $\mathcal{L}$ of dimension $k \leq p$ s.t. $\mathcal{L}$ is "closest" to the data.

Example: $p = 2$, $k = 1$.

What do we mean by closest?



$\Rightarrow$ that line is not the regression line!

That line is a geometrical property: the minimization of the orthogonal distance doesn't depend on the axis we've chosen, while regression depends on the chosen basis.

Let $\{\eta_1, \dots, \eta_K\}$ orthonormal basis spanning $\mathcal{L}$, so that:

$$\mathcal{L} = \text{span}\{\eta_1, \dots, \eta_K\}.$$

Problem: find $\eta_1, \dots, \eta_K$ o.c. s.t.

$$\sum_{i=1}^{n} \| (\underline{x}_i - \bar{\underline{x}}) - \sum_{j=1}^{K} \eta_j \eta_j' (\underline{x}_i - \bar{\underline{x}}) \|^2 \quad \text{is min.}$$

(d of before, call it (a).

Let $\underline{v}_i := \underline{x}_i - \bar{\underline{x}}$.

$$(a) = \| \underline{v}_i - \sum_{j=1}^{K} \eta_j \eta_j' \underline{v}_i \|^2 =$$

$$= \left(\underline{v}_i - \sum_{j=1}^{K} \eta_j \eta_j' \underline{v}_i \right)' \left(\underline{v}_i - \sum_{j=1}^{K} \eta_j \eta_j' \underline{v}_i \right) =$$

$$= \underline{v}_i' \underline{v}_i - 2 \sum_{j=1}^{K} \underline{v}_i' \eta_j \eta_j' \underline{v}_i + \left(\sum_{j=1}^{K} \eta_j \eta_j' \underline{v}_i \right)' \left(\sum_{t=1}^{K} \eta_t \eta_t' \underline{v}_i \right) =$$

$$= \underline{v}_i' \underline{v}_i - 2 \sum_{j=1}^{K} (\eta_j' \underline{v}_i)^2 + \sum_{j=1}^{K} (\eta_j' \underline{v}_i)^2 =$$

$$= \underline{v}_i' \underline{v}_i - \sum_{j=1}^{K} (\eta_j' \underline{v}_i)^2.$$

Then, $\sum_{i=1}^{n} \left(\underline{v}_i' \underline{v}_i - \sum_{j=1}^{K} (\eta_j' \underline{v}_i)^2 \right)$ is min $\iff$

$$\sum_{i=1}^{n} \sum_{j=1}^{K} (\eta_j' \underline{v}_i)^2 \quad \text{is max.}$$

$$= \sum_{j=1}^{K} \left(\sum_{i=1}^{n} \eta_j' \underline{v}_i \underline{v}_i' \eta_j \right) =$$

$$= \sum_{j=1}^{K} \eta_j' \left(\sum_{i=1}^{n} (\underline{x}_i - \bar{\underline{x}})(\underline{x}_i - \bar{\underline{x}})' \right) \eta_j =$$

$$= (n-1) S$$

$$= (n-1) \sum_{j=1}^{K} \eta_j' S \eta_j.$$

$K = 1$: $\max_{\eta : \|\eta\| = 1} \eta' S \eta = \lambda_1$, $\text{argmax} \dots = \underline{e}_1$.

By induction, $K = 2$: λ_2 and so on.

$\Rightarrow$ $\eta_1 = \underline{e}_1, \dots \eta_K = \underline{e}_K$ s.t. $S = \sum_i \lambda_i \underline{e}_i \underline{e}_i'$.

$\rightarrow \{\eta_1, \dots, \eta_K\}$ are the first K eigenvectors of S: we have found an alternative derivation of PCA!

(end)

Observation:

$$\sum_{i=1}^{n} \text{dendid}^2 \left(\underline{x}_i - \bar{\underline{x}} \,,\; \sum_{j=1}^{K} \underline{e}_j \underline{e}_j' (\underline{x}_i - \bar{\underline{x}}) \right) =$$

$$= \sum_{i=1}^{n} (\underline{x}_i - \bar{\underline{x}})' (\underline{x}_i - \bar{\underline{x}}) - (n-1) \sum_{j=1}^{K} \underline{e}_j' S \underline{e}_j = [\dots].$$

$\in \mathbb{R}$

Now,

$$\sum_{i=1}^{n} (\underline{x}_i - \bar{\underline{x}})' (\underline{x}_i - \bar{\underline{x}}) = \text{tr} \left(\sum_{i=1}^{n} (\underline{x}_i - \bar{\underline{x}})' (\underline{x}_i - \bar{\underline{x}}) \right) =$$

$$= \sum_{i=1}^{n} \text{tr} \left((\underline{x}_i - \bar{\underline{x}})' (\underline{x}_i - \bar{\underline{x}}) \right) \overset{(*)}{=}$$

$$= \text{tr} \left(\sum_{i=1}^{n} (\underline{x}_i - \bar{\underline{x}}) (\underline{x}_i - \bar{\underline{x}})' \right) =$$

$$= \text{tr}((n-1) S) = (n-1) \sum_{i=1}^{p} \lambda_i .$$

(*) Rotating property of trace:
$\text{tr}(ABC) = \text{tr}(CAB) = \text{tr}(BCA)$.

Moreover, $(n-1) \sum_{j=1}^{K} e_j' S e_j = (n-1) \sum_{i=1}^{K} \lambda_i .$

$$\Rightarrow [\dots] = (n-1) \sum_{i=1}^{p} \lambda_i - (n-1) \sum_{j=1}^{K} \lambda_j$$

$$= (n-1) \sum_{j=K+1}^{p} \lambda_j .$$

error of approximation (the dimension of the orth. basis is K, not p)

The PCA was introduced in this way by Pearson in 1900.

Extensions of PCA:

- ICA: Independent Component Analysis.
- Non-linear dimensional reduction. Very open field, we are entering the machine-learning territory.

Multivariate Gaussian Distribution (Ch. 4, JB): 24/3/20

$\underline{X} \in \mathbb{R}^p$ random vector.

$\nu_{\underline{X}} : \mathcal{B}$ (Borel sets of $\mathbb{R}^p$) $\to [0,1]$

$\nu_{\underline{X}}(B) = P(\underline{X} \in B) \quad \forall B \subseteq \mathcal{B}$.

In most cases, we can write $\nu_{\underline{X}}$ as:

$$\nu_{\underline{X}}(B) = \int_B f(\underline{x})\, d\underline{x}, \qquad f: \text{density}.$$

Properties of f:

1) $f \geq 0$.

2) $\int_{\mathbb{R}^p} f(\underline{x})\, d\underline{x} = 1$.

Definition: $\underline{X} \in \mathbb{R}^p$ random vector with density f.
Let $\underline{\mu} \in \mathbb{R}^p$, Σ $p \times p$ positive definite.
Then $\underline{X}$ is said to have Gaussian (or Normal) distribution with parameters $(\underline{\mu}, \Sigma)$, namely $\underline{X} \sim N_p(\underline{\mu}, \Sigma)$, if:

$$f(\underline{x}) = \frac{1}{\sqrt{(2\pi)^p \det \Sigma}} \exp\left(-\frac{1}{2}(\underline{x}-\underline{\mu})' \Sigma^{-1} (\underline{x}-\underline{\mu})\right).$$

probability density function of a multivariate Gaussian

$$\propto \exp\left(-\frac{1}{2} d^2_{\Sigma^{-1}}(\underline{x}, \underline{\mu})\right).$$

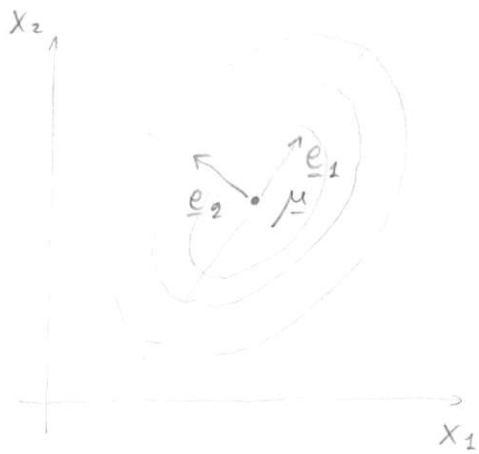
Mahalanobis distance

Contour plots for f Gaussian:

$$\{\underline{x} \in \mathbb{R}^p : f(\underline{x}) = \text{const.}\} =$$

$$= \{\underline{x} \in \mathbb{R}^p : (\underline{x}-\underline{\mu})' \Sigma^{-1} (\underline{x}-\underline{\mu}) = \text{const}^2\} =$$

$$= \{\underline{x} \in \mathbb{R}^p : d^2_{\Sigma^{-1}}(\underline{x}, \underline{\mu}) = \text{const}^2\}.$$



Writing the spectral decomposition for Σ:

$$\Sigma = \sum_{i=1}^p \lambda_i \underline{e}_i \underline{e}_i' \quad \Rightarrow \quad \Sigma^{-1} = \sum_{i=1}^p \frac{1}{\lambda_i} \underline{e}_i \underline{e}_i'.$$

Then, the axes of the ellipses will be: $\underline{e}_1, \dots, \underline{e}_p$,
while the length of the axes will be: $\propto \sqrt{\lambda_1}, \dots, \sqrt{\lambda_p}$.

Proposition 1: Let $\underline{X} \sim \mathcal{N}_p(\underline{\mu}, \Sigma)$. Then:

$$\boxed{E[\underline{X}] = \underline{\mu}}, \qquad \boxed{\text{cov}(\underline{X}) = \Sigma}.$$

Theorem 1: $\underline{X} \sim \mathcal{N}_p(\underline{\mu}, \Sigma) \iff \boxed{\underline{a}'\underline{X} \sim \mathcal{N}_1(\underline{a}'\underline{\mu}, \underline{a}'\Sigma\underline{a})}$ $\forall \underline{a} \in \mathbb{R}^p$.

This theorem gives us a characterization of the Gaussian distribution.

Proof: use the characteristic function of the Gaussian distribution. (end)

The theorem has a strong impact, both practically and theoretically.

Corollary:

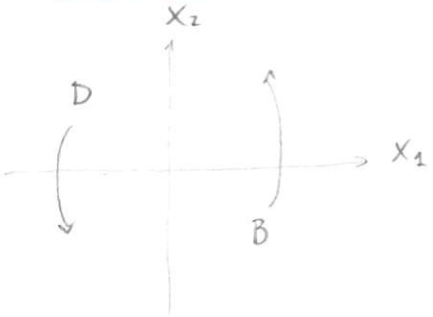
$$\underline{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}, \quad \underline{X} \sim \mathcal{N}_p(\underline{\mu}, \Sigma) \implies X_i \sim \mathcal{N}_1(\mu_i, \sigma_{ii}).$$
$$\Sigma = [\sigma_{ij}]$$

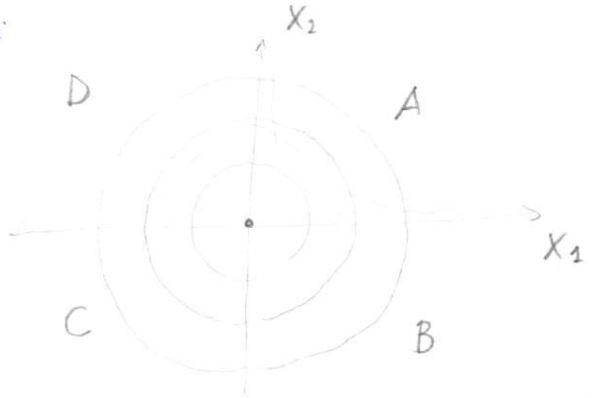
Hence, Gaussianity of a vector implies Gaussianity of its components. The vice-versa is generally wrong.

Indeed, a component is a linear combination of $\underline{X}$:

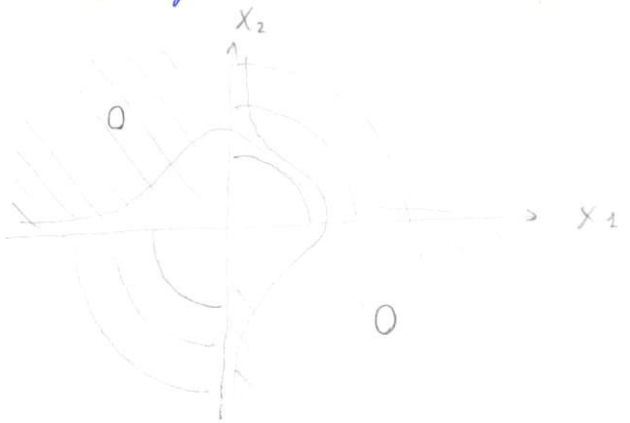
$$X_i = \underline{u}_i' \underline{X}, \quad \underline{u}_i = (0 \cdots 0 \; 1 \; 0 \cdots 0)'.$$
$$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad i$$

Example: ($\Leftarrow \not\Rightarrow$). $\underline{X} \sim \mathcal{N}_2(\underline{0}, I)$.





We get the following distribution:



This is not anymore Gaussian!

Gaussianity of components is necessary, but not sufficient!

Proposition 2: $\underline{X} \sim N_p(\underline{\mu}, \Sigma)$, A $q \times p$ matrix. Then:

$$A\underline{X} \in \mathbb{R}^q \sim N_q(A\underline{\mu}, A\Sigma A').$$

The important fact of this proposition is that Gaussianity is preserved.

Proof:

Let $\underline{a} \in \mathbb{R}^q$. We need to prove that $\underline{a}'(A\underline{X})$ has G. distr.

$$\underline{a}'(A\underline{X}) = (\underline{a}'A)\underline{X} = (A'\underline{a})'\underline{X}.$$

But $A'\underline{a} \in \mathbb{R}^p \overset{\text{(thm.)}}{\Rightarrow} (A'\underline{a})'\underline{X} \sim N_1((A'\underline{a})'\underline{\mu}, (A'\underline{a})'\Sigma(A'\underline{a}))$

$$\sim N_1(\underline{a}'(A\underline{\mu}), \underline{a}' A\Sigma A'\underline{a}) \quad \forall \underline{a} \in \mathbb{R}^q.$$

$\overset{\text{(thm1)}}{\Rightarrow} A\underline{X} \sim N_q(A\underline{\mu}, A\Sigma A')$. (end)

Proposition 3: $\underline{X} \sim N_p(\underline{\mu}, \Sigma)$, $\underline{d} \in \mathbb{R}^p$.

$$\Rightarrow \underline{X} + \underline{d} \sim N_p(\underline{\mu} + \underline{d}, \Sigma).$$

Proof: as exercise. Prove it as before. (end)

An even more interesting thing is the following. Let:

$$\underline{X} = \begin{pmatrix} \underline{X}_1 \\ \underline{X}_2 \end{pmatrix}, \quad \underline{X}_1 \in \mathbb{R}^q, \ \underline{X}_2 \in \mathbb{R}^{p-q}, \quad q < p.$$

Let $\underline{\mu} = \begin{pmatrix} \underline{\mu}_1 \\ \underline{\mu}_2 \end{pmatrix}$, $\underline{\mu}_1 \in \mathbb{R}^q$, $\underline{\mu}_2 \in \mathbb{R}^{p-q}$,

$$\Sigma = \begin{bmatrix} \Sigma_{11} \ (q \times q) & \Sigma_{12} \ (q \times (p-q)) \\ \Sigma_{21} \ ((p-q) \times q) & \Sigma_{22} \ (p \times q) \times (p-q) \end{bmatrix} \quad \text{(block notation)}$$

Proposition 4: Let $\underline{X} = \begin{pmatrix} \underline{X}_1 \\ \underline{X}_2 \end{pmatrix} \sim N_p(\underline{\mu}, \Sigma)$.

Then: $\underline{X}_1 \sim N_q(\underline{\mu}_1, \Sigma_{11})$.

Proof: Again, using linear transformations.

Let $A = [I \ (q \times q) \quad \underline{0} \ q \times (p-q)]$.

$$A\underline{X} \overset{\text{thm.}}{\sim} N_q(A\underline{\mu}, A\Sigma A').$$

But $A\underline{X} = \underline{X}_1$, $A\underline{\mu} = \underline{\mu}_1$, $A\Sigma A' = \Sigma_{11}$. (end)

Theorem 2: let $\underline{X} = \begin{pmatrix} \underline{X}_1 \\ \underline{X}_2 \end{pmatrix} \sim \mathcal{N}_p\left(\begin{pmatrix} \underline{\mu}_1 \\ \underline{\mu}_2 \end{pmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$. Then:

$$\boxed{\underline{X}_1 \perp\!\!\!\perp \underline{X}_2 \iff \Sigma_{12} = 0} \quad (\text{by symmetry, } \Sigma_{21} = 0).$$

This theorem is telling us that, in the Gaussian world, covariance equal to zero is equivalent to independence.

Proof: we can split the density of $f_{\underline{X}}$ into:

$$f_{\underline{X}} = f_{\underline{X}_1} \cdot f_{\underline{X}_2}. \qquad \text{(end)}$$

Theorem 3: let $\underline{X} = \begin{pmatrix} \underline{X}_1 \\ \underline{X}_2 \end{pmatrix} \sim \mathcal{N}_p\left(\begin{pmatrix} \underline{\mu}_1 \\ \underline{\mu}_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right)$. Then:

$$\boxed{\underline{X}_1 \mid \underline{X}_2 = x_2 \sim \mathcal{N}_q\left(\underline{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\underline{x}_2 - \underline{\mu}_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \right)}$$

(with $\det \Sigma_{22} \neq 0$).

Remark: Taking $\underline{Z}_1, \dots, \underline{Z}_p \sim$ i.i.d $\mathcal{N}_1(0,1)$, $\underline{Z} := (\underline{Z}_1, \dots, \underline{Z}_p)'$, then:

$$\underline{X} := \Sigma^{1/2}\underline{Z} + \underline{\mu} \sim \mathcal{N}_p(\underline{\mu}, \Sigma).$$

Proof (of thm 3):

Take: $A = \begin{bmatrix} I_{(q\times q)} & -\Sigma_{12}\Sigma_{22}^{-1} \; (q \times (p-q)) \\ 0 \; ((p-q)\times q) & I \quad (p-q)\times(p-q) \end{bmatrix}$.

$$A(\underline{X} - \underline{\mu}) = A\begin{pmatrix} \underline{X}_1 - \underline{\mu}_1 \\ \underline{X}_2 - \underline{\mu}_2 \end{pmatrix} = \begin{pmatrix} \underline{X}_1 - \underline{\mu}_1 - \Sigma_{12}\Sigma_{22}^{-1}(\underline{X}_2 - \underline{\mu}_2) \\ \underline{X}_2 - \underline{\mu}_2 \end{pmatrix}$$

$$\sim \mathcal{N}_p\left(\begin{pmatrix} \underline{0} \\ \underline{0} \end{pmatrix}, A\Sigma A' \right) \sim \mathcal{N}_p\left(\begin{pmatrix} \underline{0} \\ \underline{0} \end{pmatrix}, \begin{pmatrix} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & \underline{0} \\ \underline{0} & \Sigma_{22} \end{pmatrix} \right).$$

$$\underline{X}_1 - \underline{\mu}_1 - \Sigma_{12}\Sigma_{22}^{-1}(\underline{X}_2 - \underline{\mu}_2) \perp\!\!\!\perp \underline{X}_2 - \underline{\mu}_2.$$

moreover, $\underbrace{\underline{X}_1 - \underline{\mu}_1 - \Sigma_{12}\Sigma_{22}^{-1}(\underline{X}_2 - \underline{\mu}_2)}_{\underline{W}} \sim \mathcal{N}_q(\underline{0}, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$.

Now, $\underline{W} \mid \underline{X}_2 = x_2 \sim \mathcal{N}_q(\underline{0}, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$ (because of independence)
(given $\underline{X}_2 = x_2$)

$\Rightarrow \underline{X}_1 - \underline{\mu}_1 - \Sigma_{12}\Sigma_{22}^{-1}(\underline{x}_2 - \underline{\mu}_2) \sim \mathcal{N}_q(\underline{0}, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$.

By translation, $\underline{X}_1 \mid \underline{X}_2 = \underline{x}_2 \sim \mathcal{N}_q(\underline{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\underline{x}_2 - \underline{\mu}_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$. (end)

Remark: $\mathrm{Cov}(X_1 \mid X_2 = x_2) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ does not depend on x_2. → partial covariances

Example: $p = 2$.

$$\underline{X} = \begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}_2\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_{XX} & \sigma_{YX} \\ \sigma_{XY} & \sigma_{YY} \end{pmatrix}\right).$$

(X) For example, height and weight (Y).

$$Y \sim \mathcal{N}(\mu_Y, \sigma_{YY}) \Rightarrow \mathbb{P}\left[Y \in \left[\mu_Y \pm 2\sqrt{\sigma_{YY}}\right]\right] = 0.95.$$

$$Y \mid X = x \sim \mathcal{N}_1\left(\mu_Y + \sigma_{XY}\sigma_{XX}^{-1}(x - \mu_X),\ \sigma_{YY} - \sigma_{XY}\sigma_{XX}^{-1}\sigma_{XY}\right) =$$

$$\sim \mathcal{N}_1\left(\mu_Y + \frac{\sigma_{XY}}{\sigma_{XX}}(x - \mu_X),\ \sigma_{YY} - \frac{\sigma_{XY}^2}{\sigma_{XX}}\right).$$

Recall that $\rho_{XY} = \dfrac{\sigma_{XY}}{\sqrt{\sigma_{XX}\sigma_{YY}}} \in [-1, 1]$.

$$\sim \mathcal{N}_1\left(\mu_Y + \frac{\sigma_{XY}}{\sigma_{XX}}(x - \mu_X),\ \sigma_{YY}(1 - \rho^2)\right).$$

if there is some correlation between X and Y, the prediction on Y will be more accurate (less variable).

$$\boxed{y = \mathbb{E}[Y \mid X = x] = \mu_Y + \frac{\sigma_{XY}}{\sigma_{XX}}(x - \mu_X).}$$

→ linear regression function

Rmk: In the Gaussian world, the regression function is exactly linear.

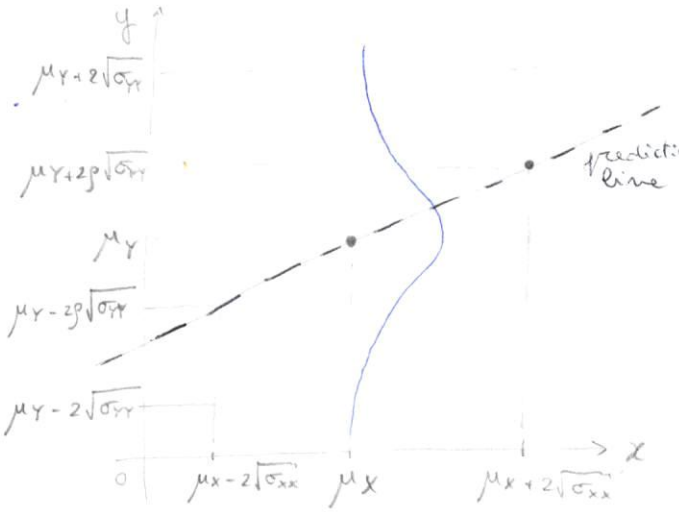
This can be rewritten as:

$$\frac{y - \mu_Y}{\sqrt{\sigma_{YY}}} = \frac{\sigma_{XY}}{\sqrt{\sigma_{YY}}\cdot\sqrt{\sigma_{XX}}} \cdot \frac{x - \mu_X}{\sqrt{\sigma_{XX}}}$$

σ_{XY}

$$\Rightarrow \frac{y - \mu_Y}{\sqrt{\sigma_{YY}}} = \rho_{XY}\frac{x - \mu_X}{\sqrt{\sigma_{XX}}}.$$

↑ standardized data



We see that an exceptionality on X (like $x = \mu_X + 2\sqrt{\sigma_{XX}}$) is not an exceptionality on Y ($y = \mu_Y + 2\rho\sqrt{\sigma_{YY}}$).

History: Sir Galton was regressing the height of fathers vs. sons.
⇒ "regressing towards mediocricy".

Regression effect $|\rho| < 1$.

The regression effect causes regression fallacy.

Example: notes on calculus I and calculus II.

Note on CI: 30 → prediction: CII 28.

Note on CI: 18 → prediction: CII 20.

The fallacy is to say: the smart guy will study less in the next exam, while the other guy was disappointed and will study more because ... ⇒ wrong reasoning.

Suggestion: Daniel Kahneman, "Thinking Fast & Slow".

Homework: read the two articles about regression fallacy. 26/3/20

Let X r.v. $\sim N_p(\mu, \Sigma)$.

Proposition: If $\det(\Sigma) > 0$, then:

$$(\underline{X}-\underline{\mu})' \Sigma^{-1} (\underline{X}-\underline{\mu}) \sim \chi^2(p).$$

Proof:

Remember: if $z_1, \dots, z_p$ are i.i.d $\sim N_1(0,1)$, then

$$\sum_{i=1}^{p} z_i^2 \sim \chi^2(p).$$

Consider $\Sigma = \sum_i \lambda_i \underline{e}_i \underline{e}_i'$ spectral decomposition,

$P = [\underline{e}_1 \dots \underline{e}_p]$, $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_p)$ ⇒ $\Sigma = P \Lambda P'$.

Consider: $\Lambda^{-\frac{1}{2}} P' (\underline{X}-\underline{\mu}) \sim N_p(\underline{0}, \Lambda^{-\frac{1}{2}} P \Sigma P' \Lambda^{-\frac{1}{2}})$.

Now, $\Lambda^{-\frac{1}{2}} P \Sigma P' \Lambda^{-\frac{1}{2}} =$

$$= \Lambda^{-\frac{1}{2}} P' P \Lambda P' P \Lambda^{-\frac{1}{2}} = \Lambda^{-\frac{1}{2}} \Lambda \Lambda^{-\frac{1}{2}} = I.$$

$\Rightarrow \underline{z} := \Lambda^{-\frac{1}{2}} P' (\underline{X}-\underline{\mu}) \sim N_p(\underline{0}, I)$.

$\underline{z} = \begin{pmatrix} z_1 \\ \vdots \\ z_p \end{pmatrix}$ ⇒ $z_1 \dots z_p$ iid $\sim N(0,1)$.

I can write:

$$(\underline{X}-\underline{\mu})' \Sigma^{-1} (\underline{X}-\underline{\mu}) = (\underline{X}-\underline{\mu})' P \Lambda^{-\frac{1}{2}} \Lambda^{-\frac{1}{2}} P' (\underline{X}-\underline{\mu}) = \underline{z}' \underline{z} =$$

$$= \sum_{i=1}^{p} z_i^2 \sim \chi^2(p).$$

(end)

Observation: if $\det(\Sigma) = 0$, let $K = \text{rank}(\Sigma)$.

$$\Sigma = \sum_{i=1}^{p} \lambda_i \underline{e}_i \underline{e}_i' \quad , \quad \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_K > 0 = \lambda_{K+1} = \dots = \lambda_p .$$

We can consider a "**generalized inverse**" for Σ:

$$\Sigma^- := \sum_{i=1}^{K} \frac{1}{\lambda_i} \underline{e}_i \underline{e}_i'$$

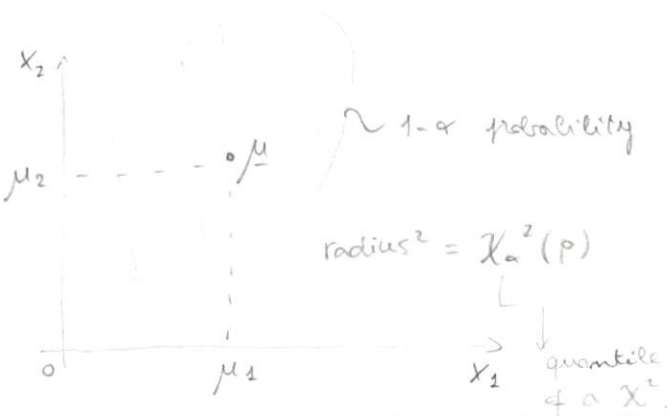
generalized inverse for non-full-rank Σ.

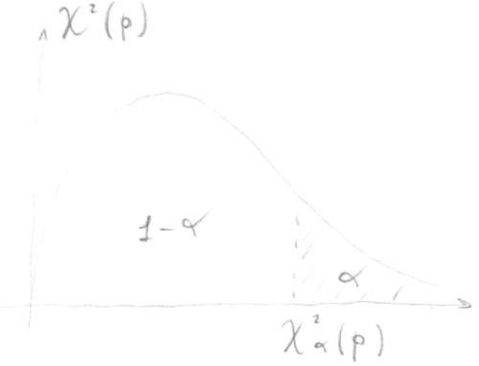
Prove that: $(\underline{X} - \underline{\mu})' \Sigma^- (\underline{X} - \underline{\mu}) \sim \chi^2(K)$.

Corollary: Let $\alpha \in (0,1)$. Then ($\text{Det}(\Sigma) > 0$):

$$Pr\left[(\underline{X} - \underline{\mu})' \Sigma^{-1} (\underline{X} - \underline{\mu}) \leq \chi^2_\alpha(p) \right] = 1 - \alpha$$

confidence region (level α) for $X \sim N_p(\mu, \Sigma)$





Remark: in R, $\chi^2_\alpha(p) = \text{qchisq}(1-\alpha, p)$. ⚠

Estimators of μ and Σ for $N_p(\mu, \Sigma)$:

$$\mathbb{X} = \begin{bmatrix} \underline{x}_1' \\ \vdots \\ \underline{x}_n' \end{bmatrix}$$ data, $\underline{x}_i$ ← obs. $\underline{X}_i$ random vector.

Hp: $\underline{X}_1, \dots, \underline{X}_n$ iid $\sim N_p(\mu, \Sigma)$, μ, Σ unknown.

Obvious choices:

- $\bar{\underline{X}} = \frac{1}{n} \sum_{i=1}^{n} \underline{X}_i$
- $S = \frac{1}{n-1} \sum_{i=1}^{n} (\underline{X}_i - \bar{\underline{X}})(\underline{X}_i - \bar{\underline{X}})'$

} unbiased estimators for μ, Σ.

MLE estimators for μ and Σ:

Suppose we want to compute the probability:

joint density $[\underline{X}_1 = \underline{x}_1, \dots, \underline{X}_n = \underline{x}_n] \overset{\text{iid}}{=} \prod_{i=1}^{n} \frac{1}{\sqrt{(2\pi)^p \text{Det}(\Sigma)}} \exp\left[-\frac{1}{2} (\underline{x}_i - \underline{\mu})' \Sigma^{-1} (\underline{x}_i - \underline{\mu}) \right]$.

Let's define now a function :

$$L : (\underline{\mu}, \Sigma) \to \text{joint density}\left[\underline{X}_1 = \underline{x}_1, \dots, \underline{X}_n = \underline{x}_n\right]$$

given $\underline{X}_1 = \underline{x}_1, \dots, \underline{X}_n = \underline{x}_n$.

L is called **likelihood function**.

Theorem: Maximum Likelihood Estimators for $\underline{\mu}, \Sigma$:

$$\underset{(\underline{\mu}, \Sigma) : \underline{\mu} \in \mathbb{R}^p, \ \Sigma \ p\times p \text{ pos. def.}}{\arg\max} L(\underline{\mu}, \Sigma \mid \underline{X}_1 = \underline{x}_1, \dots, \underline{X}_n = \underline{x}_n) = (\hat{\underline{\mu}}, \hat{\Sigma})$$

where $$L(\underline{\mu}, \Sigma \mid \underline{X}_1 = \underline{x}_1, \dots, \underline{X}_n = \underline{x}_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{(2\pi)^p \operatorname{Det}\Sigma}} \exp\left[-\frac{1}{2}(\underline{x}_i - \underline{\mu})' \Sigma^{-1} (\underline{x}_i - \underline{\mu})\right],$$

and $$\begin{cases} \hat{\underline{\mu}} = \bar{\underline{X}}, \\ \hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (\underline{X}_i - \bar{\underline{X}})'(\underline{X}_i - \bar{\underline{X}}) = \frac{n-1}{n} S^2. \end{cases}$$

MLE's for (μ, Σ)

General Properties of MLE's:

- Let $\vartheta \in \mathbb{R}^k$ parameter (ex. $\vartheta = (\underline{\mu}, \Sigma)$), $\hat{\vartheta} = \hat{\vartheta}(\text{data})$ MLE estimator of ϑ. Let $h : \mathbb{R}^k \to \mathbb{R}^j$.
  What is the MLE of $h(\vartheta)$? This is written $\widehat{h(\vartheta)}$.
  The **invariance property** of MLE holds: $\widehat{h(\vartheta)} = h(\hat{\vartheta})$.

Example: $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (\underline{X}_i - \bar{\underline{X}})'(\underline{X}_i - \bar{\underline{X}})$, MLE for Σ

if $\underline{X}_1, \dots, \underline{X}_n$ iid $\sim N_p(\underline{\mu}, \Sigma)$.

Which is a good estimator of λ_1, where $\Sigma = \sum_i \lambda_i \underline{e}_i \underline{e}_i'$?

By the property of before, $\hat{\lambda}_1$ is s.t. $\hat{\Sigma} = \sum_i \hat{\lambda}_i \hat{\underline{e}}_i \hat{\underline{e}}_i'$.

Sampling distribution of $\bar{\underline{X}}$ and S^2 (or $\hat{\Sigma}$):

Assume $\underline{X}_1, \dots, \underline{X}_n$ iid $\sim N_p(\underline{\mu}, \Sigma)$.

Proposition: $\bar{\underline{X}} \sim N_p\left(\underline{\mu}, \frac{\Sigma}{n}\right)$

distribution of $\bar{\underline{X}} = \frac{1}{n} \sum_{i=1}^{n} \underline{X}_i$

Proof: $\tilde{\underline{X}} = \begin{pmatrix} \underline{X}_1 \\ \underline{X}_2 \\ \vdots \\ \underline{X}_n \end{pmatrix} \in \mathbb{R}^{np}$ $np \times 1$ matrix.

Then (iid) $$\tilde{\underline{X}} \sim N_{np}\left(\begin{pmatrix} \underline{\mu} \\ \underline{\mu} \\ \vdots \\ \underline{\mu} \end{pmatrix}, \begin{pmatrix} \Sigma & & 0 \\ & \Sigma & \\ 0 & & \Sigma \end{pmatrix}\right).$$

Consider $A = [\underbrace{I_{(p\times p)}\ I\ \dots\ I}_{n}]$ $\quad p \times np$.

$\Rightarrow \frac{1}{n} A \tilde{X} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Indeed, $A = \begin{bmatrix} 1 \cdots 1 & 1 \cdots 1 & & 1 \cdots 1 \\ & \ddots & \cdots & \ddots \\ & 1 & & 1 \end{bmatrix}$

$\Rightarrow \frac{1}{n} A \tilde{X} = \frac{1}{n} \begin{pmatrix} \sum_{i=1}^{n} X_{i1} \\ \vdots \\ \sum_{i=1}^{n} X_{ip} \end{pmatrix} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Hence: $\bar{X} = \frac{1}{n} A \tilde{X} \sim N_p\left(\frac{1}{n} A \begin{pmatrix} \mu \\ \vdots \\ \mu \end{pmatrix}, \frac{1}{n^2} A \begin{pmatrix} \Sigma & & 0 \\ & \ddots & \\ 0 & & \Sigma \end{pmatrix} A'\right)$

$\sim N_p\left(\mu, \frac{1}{n^2} n\Sigma\right) \sim N_p\left(\mu, \frac{1}{n}\Sigma\right)$.

(end)

▶ What about the distribution of S? We are talking about a random positive-semi definite matrix.

Definition: Let $Z_1, \dots, Z_m$ iid $\sim N_p(0, \Sigma)$, $\det(\Sigma) > 0$.

Then $\sum_{i=1}^{m} Z_i Z_i' \sim \text{Wishart}(\Sigma, m)$

Wishart Distribution

m: degrees of freedom

Properties of Wishart distribution:

1) $A_1 \sim \text{Wish}(\Sigma, m_1)$, $A_2 \sim \text{Wish}(\Sigma, m_2)$, $A_1 \perp\!\!\!\perp A_2$ (stoch. independent)

$\Rightarrow A_1 + A_2 \sim \text{Wish}(\Sigma, m_1 + m_2)$.

sums of Wisharts

Proof:

$$\begin{cases} A_1 = \sum_{i=1}^{m_1} Z_i Z_i', & Z_1, \dots, Z_{m_1} \text{ iid} \sim N_p(0, \Sigma) \\ A_2 = \sum_{i=1}^{m_2} \tilde{Z}_i \tilde{Z}_i', & \tilde{Z}_1, \dots, \tilde{Z}_{m_2} \text{ iid} \sim N_p(0, \Sigma) \end{cases}$$

Renaming the rvs:

$Z_1, \dots, Z_{m_1}, \tilde{Z}_1, \dots, \tilde{Z}_{m_2}$

$\quad\|\quad\ \|\quad\quad\ \|\quad\quad\ \|\quad\quad\ \|$

$W_1 \quad W_2 \quad\ W_{m_1} \quad W_{m_1+1} \quad W_{m_1+m_2}$

$W_1, \dots, W_{m_1+m_2}$ iid $\sim N_p(0, \Sigma)$.

$\Rightarrow A_1 + A_2 = \sum_{i=1}^{m_1} Z_i Z_i' + \sum_{i=1}^{m_2} \tilde{Z}_i \tilde{Z}_i' = \sum_{i=1}^{m_1+m_2} W_i W_i' \sim \text{Wish}(\Sigma, m_1+m_2)$.

(end)

2) Let C $k \times p$ constant, $A \sim \text{Wish}(\Sigma, m)$.

$\Rightarrow$ $\boxed{CAC' \sim \text{Wish}(C\Sigma C', m).}$

Proof: $A = \sum_{i=1}^{m} \underline{z}_i \underline{z}_i'$, $\underline{z}_i$ iid $\sim \mathcal{N}_p(\underline{0}, \Sigma)$.

$CAC' = \sum_{i=1}^{m} C\underline{z}_i \underline{z}_i' C'$, $\underline{W}_i$ iid $\sim \mathcal{N}_k(\underline{0}, C\Sigma C')$.

$\underline{W}_i$ $\sim \text{Wish}(C\Sigma C', m)$ by definition.

(end)

3) $\sigma^2 > 0$, $A \sim \text{Wish}(\Sigma, m)$.

$\Rightarrow$ $\boxed{\sigma^2 A \sim \text{Wish}(\sigma^2 \Sigma, m)}$

Proof: $\sigma^2 A = \sum_{i=1}^{m} \sigma \underline{z}_i \underline{z}_i' \sigma$

$\underline{W}_i \Rightarrow \underline{W}_i$ iid $\sim \mathcal{N}_p(\underline{0}, \sigma^2\Sigma)$.

$\sim \text{Wish}(\sigma^2\Sigma, m)$ by def. (end)

4) Let $A \sim \text{Wish}(\Sigma, m)$ s.t. $\boxed{\Sigma \in \mathbb{R}}$ (i.e. $p = 1$), $(\Sigma = \sigma^2)$

Then: $A = \sum_{i=1}^{m} z_i z_i$, $z_1, \dots, z_m$ iid $\sim \mathcal{N}_1(0, \underbrace{\Sigma}_{\sigma^2})$.

$\Rightarrow \frac{1}{\sigma^2} A = \sum_{i=1}^{m} \frac{z_i}{\sigma} \cdot \frac{z_i}{\sigma}$ $\Rightarrow \frac{1}{\sigma^2} A \sim \chi^2(m)$.

W_i iid $\mathcal{N}_1(0,1)$

In conclusion,

$A \sim \text{Wish}(\Sigma, m)$, $\Sigma \in \mathbb{R}$.

$\Rightarrow$ $\boxed{A \sim \Sigma \cdot \chi^2(m).}$, i.e. $\frac{A}{\Sigma} \sim \chi^2(m)$. Wishart in scalar case

Exercise: Let $\underline{c} \in \mathbb{R}^p$, $A \sim \text{Wish}(\Sigma, m)$, $\Sigma \in \mathbb{R}^{p \times p}$.

$\underline{c}' A \underline{c} \sim ?$

$\underline{c}' A \underline{c} \overset{2)}{\sim} \text{Wish}(\underline{c}' \Sigma \underline{c}, m)$.

But now, $\underline{c}' A \underline{c} > 0$ because Σ positive definite.

$\Rightarrow \text{Wish}(\underline{c}' \Sigma \underline{c}, m) \sim (\underline{c}' \Sigma \underline{c}) \chi^2(m)$.

$\Rightarrow$ $\boxed{\underline{c}' A \underline{c} \sim \underline{c}' \Sigma \underline{c} \, \chi^2(m).}$

Theorem: $\underline{X}_1, \dots, \underline{X}_n$ iid $\sim N_p(\underline{\mu}, \Sigma)$.

$$\Rightarrow \quad \sum_{i=1}^{n} (\underline{X}_i - \bar{\underline{X}})(\underline{X}_i - \bar{\underline{X}})' \sim \text{Wish}(\Sigma, n-1).$$

Corollary:

$$S' = \frac{1}{n-1} \sum_{i=1}^{n} (\underline{X}_i - \bar{\underline{X}})(\underline{X}_i - \bar{\underline{X}}) \sim \text{Wish}\left(\frac{1}{n-1}\Sigma, n-1\right).$$

distribut. of S

$$\hat{\Sigma} \sim \text{Wish}\left(\frac{1}{n}\Sigma, n-1\right).$$

We'll summarize all the informations about the estimators in a single theorem:

Theorem: $\underline{X}_1, \dots, \underline{X}_n$ iid $\sim N_p(\underline{\mu}, \Sigma)$. Then:

1) $\bar{\underline{X}} \sim N_p(\underline{\mu}, \frac{1}{n}\Sigma)$,
2) $(n-1)S' \sim \text{Wish}(\Sigma, n-1)$,
3) $\bar{\underline{X}} \perp\!\!\!\perp S'$ (stoch. independent).

distributions of $\underline{X}, S$ in the multivariate Gaussian setting

Indeed, we have another theorem:

Theorem: $\bar{\underline{X}}$ and S' are sufficient statistics. [?]

▷ Useful transformations for making data more Gaussian:

If X is a proportion $\in [a, 1]$, a useful transformation is:

$$\log \frac{x}{1-x} =: \text{logit}(x)$$

Asymptotic results for (Gaussian) r.vs:

1) LLN: (Law of Large Numbers)

$\underline{X}_1, \dots, \underline{X}_n$, random vectors iid s.t. $E[\underline{X}_i] = \underline{\mu}$, $\text{cov}(\underline{X}_i) = \Sigma$ exist. Then:

$$\begin{cases} \bar{\underline{X}} = \frac{1}{n}\sum_{i=1}^{n} \underline{X}_i \xrightarrow{P} \underline{\mu} & \text{as } n \to \infty. \\ S = \frac{1}{n-1}\sum_{i=1}^{n} (\underline{X}_i - \bar{\underline{X}})(\underline{X}_i - \bar{\underline{X}})' \xrightarrow{P} \Sigma & \text{as } n \to \infty. \end{cases}$$

2) CLT: (Teorema Centrale del Limite)

$\underline{X}_1, \dots, \underline{X}_n, \dots$ r.v. iid s.t. $E[\underline{X}_i] = \underline{\mu}$, $Cov(\underline{X}_i) = \Sigma$ exist,

then: $\boxed{\sqrt{n}(\bar{\underline{X}} - \underline{\mu}) \sim AN_p(\underline{0}, \Sigma)}$.

↳ asymptotically normal

Meaning: for large n, one can approximate the distribution of $\sqrt{n}(\bar{\underline{X}} - \underline{\mu})$ with a $N_p(\underline{0}, \Sigma)$.

In practice: for large n, $\bar{\underline{X}} \sim N_p(\underline{\mu}, \frac{1}{n}\Sigma)$ approximately.

27/3/20

Lab Session:

Lab 2: quick explanation.
↳ Look at the code yourself.

Lab 3: PCA.

Inference for the mean $\underline{\mu}$: 30/3/20

We want to use the partial information we have in order to infer the data of the population.

$\mathbb{R}^p \ni \underline{X}_1, \dots, \underline{X}_n$ r. vectors iid, with $E[\underline{X}_i] = \underline{\mu}$ and $Cov(\underline{X}_i) = \Sigma$, $\det \Sigma > 0$.

$\bar{\underline{X}} = \frac{1}{n}\sum_{i=1}^{n}$ (point) estimator for $\underline{\mu}$.

I) Large n: $(n >> p)$

i) $\sqrt{n}(\bar{\underline{X}} - \underline{\mu}) \underset{\text{approx.}}{\sim} N_p(\underline{0}, \Sigma)$ (CLT)

$\Rightarrow (\sqrt{n}(\bar{\underline{X}} - \underline{\mu}))' \Sigma^{-1} (\sqrt{n}(\bar{\underline{X}} - \underline{\mu})) \sim \chi^2(p)$,

or $n(\bar{\underline{X}} - \underline{\mu})' \Sigma^{-1} (\bar{\underline{X}} - \underline{\mu}) \sim \chi^2(p)$.

The problem is that we don't know Σ as well.
If Σ is unknown, we have that:

ii) $S \xrightarrow{P} \Sigma$ as $n \to \infty$. (LLN)

Hence, for large n,

$$n(\bar{\underline{X}} - \underline{\mu})' S^{-1} (\bar{\underline{X}} - \underline{\mu}) \sim \chi^2(p).$$

distribution of the pivotal statistics with large n

This is called a pivotal statistics: it's not a quantity that we know, but its distribution is known; we can make inference on it!

Consequences:

$$\Pr\left[n(\bar{\underline{X}} - \underline{\mu})' S^{-1} (\bar{\underline{X}} - \underline{\mu}) \leq \chi^2_\alpha(p) \right] = 1 - \alpha \quad \text{if } \alpha \in (0,1).$$

$$\Rightarrow \Pr\left[d^2_{S^{-1}}(\bar{\underline{X}}, \underline{\mu}) \leq \chi^2_\alpha(p) \right] = 1 - \alpha. \qquad (*)$$

Define now the following sets:

$$\mathcal{E}^\alpha_{S^{-1}}(\underline{\mu}) := \left\{ \underline{x} \in \mathbb{R}^p : d^2_{S^{-1}}(\bar{\underline{X}}, \underline{\mu}) \leq \chi^2_\alpha(p) \right\}$$

$$\mathcal{E}^\alpha_{S^{-1}}(\bar{\underline{X}}) := \left\{ \underline{\eta} \in \mathbb{R}^p : d^2_{S^{-1}}(\underline{\eta}, \bar{\underline{X}}) \leq \chi^2_\alpha(p) \right\}$$

Two ellipses, same axes, same radius. One centered on $\underline{\mu}$, the other on $\bar{\underline{X}}$. Their relationship is:

$$\bar{\underline{X}} \in \mathcal{E}^\alpha_{S^{-1}}(\underline{\mu}) \iff \underline{\mu} \in \mathcal{E}^\alpha_{S^{-1}}(\bar{\underline{X}}).$$

So $(*)$ is equivalent to:

$$\Pr\left[\bar{\underline{X}} \in \mathcal{E}^\alpha_{S^{-1}}(\underline{\mu}) \right] = 1 - \alpha.$$

($\bar{\underline{X}}$: random part)

$$\Updownarrow$$

$$\Pr\left[\underline{\mu} \in \mathcal{E}^\alpha_{S^{-1}}(\bar{\underline{X}}) \right] = 1 - \alpha.$$

(random part)

- Confidence region (level α) for $\underline{\mu}$:

$$CR_{1-\alpha}(\underline{\mu}) = \left\{ \underline{\eta} \in \mathbb{R}^p : \underline{\eta} \in \mathcal{E}^\alpha_{S^{-1}}(\bar{\underline{X}}) \right\} =$$

$$= \left\{ \underline{\eta} \in \mathbb{R}^p : n(\underline{\eta} - \bar{\underline{X}})' S^{-1} (\underline{\eta} - \bar{\underline{X}}) \leq \chi^2_\alpha(p) \right\}.$$

Basics of inferential statistics:

Test: $H_0 : \underline{\mu} = \underline{\mu}_0$ vs. $H_1 : \underline{\mu} \neq \underline{\mu}_0$.

$\alpha \in (0,1)$ (small): level of confidence of the test.

Consider the statistics:

$$T_0^2 := n\,(\bar{\underline{X}} - \underline{\mu}_0)'\, S^{-1}\, (\bar{\underline{X}} - \underline{\mu}_0).$$

If H_0 is true, then $\boxed{T_0^2 \overset{(H_0)}{\sim} \chi^2(p).}$

$\Rightarrow$ We have proof against H_0 if $\bar{\underline{X}}$ is very far from $\underline{\mu}_0$.

How big should that distance be?

Reject if $T_0^2 > \chi^2_\alpha(p)$. Either is wrong the assumption, or we are observing a "miracle" (something that happens α times every 1).

$$\text{Rejection Region}_\alpha = \left\{ T_0^2 > \chi^2_\alpha(p) \right\}.$$

rejection region (level α) for H_0

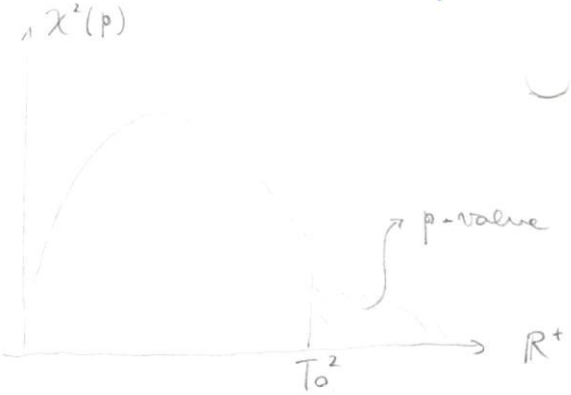
Normally we don't fix α, but instead we look for the p-value of the test.

Definition:

p-value: area of $\chi^2(p)$ at the right of T_0^2.

Testing with p-value: from the definition,

$$\text{p-value} \leq \alpha \iff T_0^2 > \chi^2_\alpha(p).$$



- p-value small: we can reject H_0.
- p-value large: we cannot reject H_0.

History of 5% - 1%: values computed by Fisher on the tables.
↳ only a "historical reason". There is no absolute definition of what is large and small.

All this inference is based on knowing the distribution of the pivotal statistics. What if the sample is not large enough? We want to build another pivotal statistics for small n.

II) Small n:

Fisher's idea: convert the variables into Gaussians.

- 1905: "Gosset", brewer of the Guinness factory in Dublin, aka "Student" → father of the "Student-t".

Key Assumption: $\boxed{\underline{X}_1, \dots, \underline{X}_n \text{ iid} \sim N_p(\underline{\mu}, \Sigma)}$, $\det \Sigma > 0$.
↳ key fact

What is the distribution of

$$n\,(\bar{\underline{X}} - \underline{\mu})'\, S^{-1}\, (\bar{\underline{X}} - \underline{\mu}) \; ?$$

We need to introduce Fisher's distribution.

Definition: let $Y \sim \chi^2(n)$, $W \sim \chi^2(m)$, $Y \perp\!\!\!\perp W$.

$$\Rightarrow \quad \frac{Y/n}{W/m} \sim F(n, m).$$

Fisher's Distribution

↳ coming from Fisher

Observations: properties of Fisher's distributions.

1) $t \sim t(n) \quad \Rightarrow \quad t = \dfrac{Z}{\sqrt{\frac{W}{n}}}$, $\begin{cases} Z \sim \mathcal{N}(0,1) \\ W \sim \chi^2(n) \\ Z \perp\!\!\!\perp W \end{cases}$.

Taking:

$$t^2 = \frac{Z^2}{W/n}, \quad \begin{cases} Z^2 \sim \chi^2(1) \\ W \sim \chi^2(n) \\ Z^2 \perp\!\!\!\perp W \end{cases} \Rightarrow \quad t^2 \sim F(1, m).$$

Fisher with $n=1$ d.o.f.

↳ squared of a t-Student distribution.

2) $F(n,m) \xrightarrow{m \to \infty} \frac{1}{n}\chi^2(n)$ (convergence in distribution)

Fisher with $m \to \infty$ d.o.f's

Indeed, $\dfrac{Y/n}{W/m}$, $Y \sim \chi^2(n)$, $W \sim \chi^2(m)$, $Y \perp\!\!\!\perp W$.

$$\Rightarrow \quad W = \sum_{i=1}^{m} Z_i^2, \quad Z_1, \dots, Z_m \text{ iid } \sim \mathcal{N}(0,1).$$

$$\mathbb{E}[Z_i^2] = 1 \quad \Rightarrow \quad \frac{1}{m}\sum_{i=1}^{m} Z_i^2 \to 1 \text{ as } m \to \infty \quad \text{(LLN)}$$

$$\Rightarrow \quad \frac{W}{m} \to 1 \text{ as } m \to \infty \quad \Rightarrow \quad \frac{Y/n}{W/m} \to \frac{1}{n}Y \sim \frac{1}{n}\chi^2(n) \text{ as } m \to \infty.$$

Exercise: use R to make plots of different F distributions.

Theorem: Hotelling, 1931.

Suppose $\underline{X} \sim \mathcal{N}_p(\underline{\mu}, \Sigma)$, $\det \Sigma > 0$, $W \sim \text{Wish}\left(\frac{1}{m}\Sigma, m\right)$, Σ $p \times p$, $\underline{X} \perp\!\!\!\perp W$. Then:

$$\frac{m-p+1}{mp}(\underline{X}-\underline{\mu})' W^{-1}(\underline{X}-\underline{\mu}) \sim F(p, m-p+1).$$

Hotelling's Theorem

Corollary: let $\underline{X}_1, \dots, \underline{X}_n$ iid $\sim \mathcal{N}_p(\underline{\mu}, \Sigma)$.

$$\Rightarrow \quad n(\bar{\underline{X}}-\underline{\mu})' S^{-1}(\bar{\underline{X}}-\underline{\mu}) \sim \frac{(n-1)p}{n-p} F(p, n-p).$$

distribution of pivotal statistics with Gaussianity

Proof (corollary):

$\sqrt{n}(\bar{\underline{X}} - \underline{\mu}) \sim \mathcal{N}_p(\underline{0}, \Sigma)$, $S \sim \text{Wish}\left(\frac{1}{n-1}\Sigma, n-1\right)$, $\bar{\underline{X}} \perp\!\!\!\perp S$.

Use Hotelling's thm. $(m = n-1)$:

$$\frac{n-1+p+1}{(n-1)p}\, n\,(\bar{\underline{X}} - \underline{\mu})'\, S^{-1}\,(\bar{\underline{X}} - \underline{\mu}) \sim F(p, n-1-p+1)$$

$$\Rightarrow \quad \frac{n-p}{(n-1)p}\, n\,(\bar{\underline{X}} - \underline{\mu})'\, S^{-1}\,(\bar{\underline{X}} - \underline{\mu}) \sim F(p, n-p)$$

$$\Leftrightarrow \quad n\,(\bar{\underline{X}} - \underline{\mu})'\, S^{-1}\,(\bar{\underline{X}} - \underline{\mu}) \sim \frac{(n-1)p}{n-p} F(p, n-p)$$

(end)

We have discovered the distribution of $T^2 := n\,(\bar{\underline{X}} - \underline{\mu})'\, S^{-1}\,(\bar{\underline{X}} - \underline{\mu})$! This is the pivotal statistics for making inference with $\underline{\mu}$.

- **Confidence Region:**

$$Pr\Big[\underbrace{n\,(\bar{\underline{X}} - \underline{\mu})'\, S^{-1}\,(\bar{\underline{X}} - \underline{\mu})}_{d^2_{S^{-1}}(\bar{\underline{X}}, \underline{\mu})} \leq \frac{(n-1)p}{n-p} F_\alpha(p, n-p)\Big] = 1-\alpha$$

$$\Rightarrow \quad Pr\left[\underline{\mu} \in \mathcal{E}^\alpha_{S^{-1}}(\bar{\underline{X}})\right] = 1-\alpha \quad , \text{ with } \alpha \in (0,1) \text{ and}$$

$$\mathcal{E}^\alpha_{S^{-1}}(\bar{\underline{X}}) = \Big\{\underline{\eta} \in \mathbb{R}^p : d^2_{S^{-1}}(\bar{\underline{X}}, \underline{\mu}) \leq \underbrace{\frac{(n-1)p}{n-p} F_\alpha(p, n-p)}_{\text{same Mahalanobis distance as before, but different radius!}}\Big\}.$$

$$\Rightarrow \quad \boxed{CR_{1-\alpha}(\underline{\mu}) = \mathcal{E}^\alpha_{S^{-1}}(\bar{\underline{X}})}.$$

Observation:

$$\boxed{\frac{(n-1)p}{n-p} F_\alpha(p, n-p) \to \chi^2_\alpha(p)} \quad \text{as } n \to \infty.$$

$\downarrow 1$ $\qquad$ $\to \frac{1}{p}\chi^2(p)$

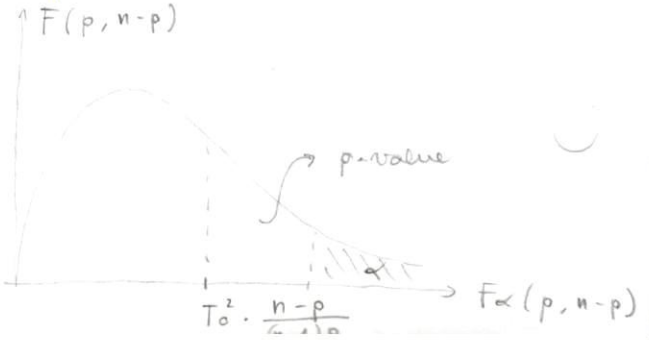
convergence of the two pivotal quantities as $n \to \infty$

- **Testing:** $H_0 : \underline{\mu} = \underline{\mu}_0$ vs. $H_1 : \underline{\mu} \neq \underline{\mu}_0$ at level $\alpha \in (0,1)$.

$$T_0^2 = n\,(\bar{\underline{X}} - \underline{\mu}_0)'\, S^{-1}\,(\bar{\underline{X}} - \underline{\mu}_0).$$

If H_0 is true $\Rightarrow$ $T_0^2 \overset{(H_0)}{\sim} \frac{(n-1)p}{n-p} F(p, n-p)$.

Reject if $T_0^2 > \frac{(n-1)p}{n-p} F_\alpha(p, n-p)$.



Remark: this time, the p-value is the area at the right of:

$T_0^2 \cdot \frac{n-p}{(n-1)p}$. Don't forget the rescaling factor!

Example: $\underline{X}_1, \dots, \underline{X}_{10}$ iid $\sim N_2(\underline{\mu}, \Sigma)$.

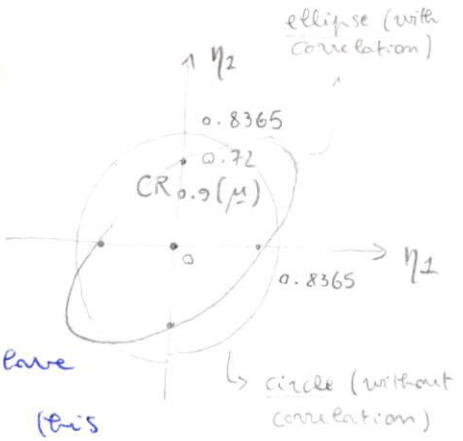
Suppose that $\bar{\underline{X}} = \underline{0}$, and $S = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.

$$CR_{1-\alpha}(\underline{\mu}) = \left\{ \underline{\eta} \in \mathbb{R}^p : 10\, \underline{\eta}' I \underline{\eta} \leq \frac{9 \cdot 2}{8} F_\alpha(2,8) \right\}.$$

Choose $\alpha = 10\,\%$ $\Rightarrow$ $\begin{cases} 1-\alpha = 0.9. \\ F_{0.1}(2,8) = 3.11. \end{cases}$

$$\Rightarrow CR_{0.9}(\underline{\mu}) = \left\{ \underline{\eta} \in \mathbb{R}^p : 10(\eta_1^2 + \eta_2^2) \leq 6.997 \right\}$$

$$= \left\{ \underline{\eta} \in \mathbb{R}^2 : \eta_1^2 + \eta_2^2 \leq 0.6997 \right\}.$$



Observe that $\underline{\mu}$ is a parameter, and what is random is the circle we have drawn. We have tossed the coin and today we have observed this ellipse. This ellipse is right 90 % of the times.

Assume now $\bar{\underline{X}} = \underline{0}$, $S = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$.

$$\Rightarrow CR_{0.9}(\underline{\mu}) = \left\{ \underline{\eta} \in \mathbb{R}^2 : 10\, \underline{\eta}' S^{-1} \underline{\eta} \leq \frac{9 \cdot 2}{8} \cdot 3.11 \right\}$$

$$S^{-1} = \begin{bmatrix} 4/3 & -2/3 \\ -\frac{2}{3} & 4/3 \end{bmatrix} \Rightarrow 10\, \underline{\eta}' S^{-1} \underline{\eta} = \frac{10 \cdot 4}{3} \left(\eta_1^2 - \eta_1 \eta_2 + \eta_2^2 \right).$$

$$\Rightarrow CR_{0.9}(\underline{\mu}) = \left\{ \underline{\eta} \in \mathbb{R}^2 : \eta_1^2 - \eta_1 \eta_2 + \eta_2^2 \leq \frac{3}{4} \cdot 0.6997 \right\}.$$

↳ we have an ellipse this time! It changed shape because Σ changed. The more the two components are correlated, the more the two variables will be squished to a bisector.

- **univariate vs. multivariate CIs:**

Now, $\bar{\underline{X}} = \begin{pmatrix} \bar{X}_1 \\ \bar{X}_2 \end{pmatrix} = \underline{0}$, $S = \begin{pmatrix} s_{11} = 1 & s_{12} \\ s_{21} & s_{22} = 1 \end{pmatrix}$.

Consider the univariate confidence intervals:

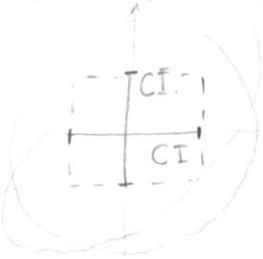
$$\begin{cases} C.I_{0.9}(\mu_1) = \left[\bar{X}_1 \pm t_{0.95}(9) \sqrt{\frac{1}{10}} \right] = \left[\pm 0.5796 \right] \\ CI_{0.9}(\mu_2) = \left[\pm 0.5796 \right] \end{cases}$$

Here we didn't use the covariance for these Confidence Intervals!

CI's are independent from S.

They seem better intervals than before!

But here we have something fishy:



$$\Pr\left[\mu_1 \in CI_{0.9}(\bar{X}_1),\ \mu_2 \in CI_{0.9}(\bar{X}_2)\right] = 0.81$$

$$\overset{\bar{X}_1 \perp\!\!\!\perp \bar{X}_2}{=} \Pr\left[\mu_1 \in CI_{0.9}(\bar{X}_1)\right] \Pr\left[\mu_2 \in CI_{0.9}(\bar{X}_2)\right] = (0.9)^2 < 0.9.$$

Hence, the probability for (μ_1, μ_2) to fall inside the cartesian square is actually less than 0.9! In other words, we are giving a multivariate CI with a level of $(1-\alpha)^2$, smaller: the true α here is $\alpha = 19\%$.

31/3/20

Quick presentation: 7th of april.

First slide: title of the dataset.
Second slide: faces.

Description of the dataset, what we have already done.

Suggestion of the professor: follow the pace of the professor's notes, but study on the book. Don't study just the notes!

Confidence Intervals for linear combinations of the mean:

We are still in the small n sample:

$$\underline{X}_1, \dots, \underline{X}_n \ \text{iid} \sim \mathcal{N}_p(\underline{\mu}, \Sigma)$$

$$\bar{\underline{X}} = \frac{1}{n}\sum_{i=1}^{n} \underline{X}_i \quad \text{estimator for } \underline{\mu}.$$

I) • One-at-the-time CI for μ:

Let $\underline{a} \in \mathbb{R}$. What is $CI_{1-\alpha}(\underline{a}'\underline{\mu})$?

Estimator for $\underline{a}'\underline{\mu}$: $\underline{a}'\bar{\underline{X}}$.

$$\underline{a}'\bar{\underline{X}} \sim \mathcal{N}_1\left(\underline{a}'\underline{\mu}, \frac{1}{n}\underline{a}'\Sigma\underline{a}\right), \quad \text{i.e.} \quad \frac{\underline{a}'\bar{\underline{X}} - \underline{a}'\underline{\mu}}{\sqrt{\underline{a}'\Sigma\underline{a}}}\sqrt{n} \sim \mathcal{N}_1(0,1).$$

We don't know Σ, but we can estimate it with S:

$$(n-1)S \sim \text{Wishart}(\Sigma, n-1)$$

$$\Rightarrow (n-1)\,\underline{a}'S\underline{a} \sim (\underline{a}'\Sigma\underline{a})\,\chi^2(n-1).$$

Moreover, $\bar{\underline{X}} \perp\!\!\!\perp S$. Taking a ratio between a Gaussian and a normalized squared Wishart (χ), we get a Student-t:
(by definition of Student-t)

$$\frac{\dfrac{\underline{a}'\bar{\underline{X}} - \underline{a}'\underline{\mu}}{\sqrt{\underline{a}'\Sigma\underline{a}}}\sqrt{n} \sim \mathcal{N}(0,1)}{\sqrt{\dfrac{(n-1)\,\underline{a}'S\underline{a}}{(n-1)(\underline{a}'\Sigma\underline{a})}} \sim \sqrt{\dfrac{\chi^2(n-1)}{n-1}}} \sim t(n-1).$$

Simplifying:

$$\frac{\underline{a}'\bar{\underline{X}} - \underline{a}'\underline{\mu}}{\sqrt{\underline{a}'S\underline{a}}}\sqrt{n} \sim t(n-1).$$

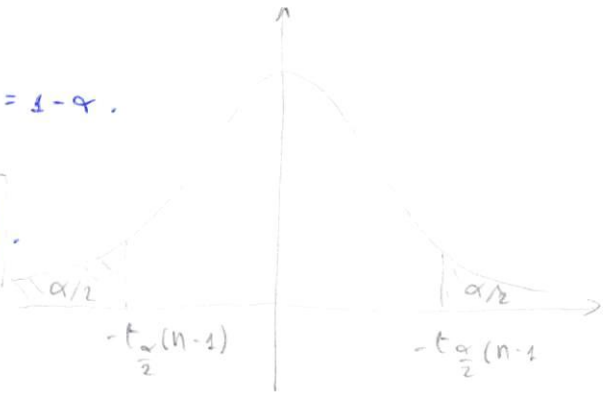
pivotal quantity for one-at-the-time CI of $\underline{a}'\underline{\mu}$

We have obtained our pivotal quantity! For example,

$$\Pr\left[\frac{|\underline{a}'\bar{\underline{X}} - \underline{a}'\underline{\mu}|}{\sqrt{\underline{a}'S\underline{a}}}\sqrt{n} < t_{\frac{\alpha}{2}}(n-1)\right] = 1-\alpha, \quad \alpha \in (0,1).$$

$$\Rightarrow \Pr\left[\underline{a}'\underline{\mu} \in \left[\underline{a}'\bar{\underline{X}} \pm t_{\frac{\alpha}{2}}(n-1)\sqrt{\frac{\underline{a}'S\underline{a}}{n}}\right]\right] = 1-\alpha.$$

$$\Rightarrow CI_{1-\alpha}(\underline{a}'\underline{\mu}) = \left[\underline{a}'\bar{\underline{X}} \pm t_{\frac{\alpha}{2}}(n-1)\sqrt{\frac{\underline{a}'S\underline{a}}{n}}\right].$$



Example: $\underline{a} = (0 \cdots 0\ \underset{i}{1}\ 0 \cdots 0)'$, $i = 1 \ldots p$.

$$\Rightarrow \underline{a}'\underline{\mu} = \mu_i.$$

$$CI_{1-\alpha}(\mu_i) = \left[\bar{X}_i \pm t_{\frac{\alpha}{2}}(n-1)\sqrt{\frac{S_{ii}}{n}}\right].$$

If $\underline{a} = (0 \cdots 0\ \underset{i}{1}\ 0 \cdots 0\ \underset{j}{-1}\ 0 \cdots 0)'$ $\Rightarrow \underline{a}'\underline{\mu} = \mu_i - \mu_j$.

$$\Rightarrow CI_{1-\alpha}(\mu_i - \mu_j) = \left[\bar{X}_i - \bar{X}_j \pm t_{\frac{\alpha}{2}}(n-1)\sqrt{\frac{S_{ii} - 2S_{ij} + S_{jj}}{n}}\right]. \quad \square$$

We proved that:

$$\Pr\left[\underline{a}'\underline{\mu} \in \left[\underline{a}'\bar{\underline{X}} \pm t_{\frac{\alpha}{2}}(n-1)\sqrt{\frac{\underline{a}'S\underline{a}}{n}}\right]\right] = 1-\alpha, \quad \forall \underline{a} \in \mathbb{R}^p.$$

one-at-the-time CI for $\underline{a}'\underline{\mu}$

Aside: testing $H_0: \underline{a}'\underline{\mu} = \delta_0$.

1) $H_1: \underline{a}'\underline{\mu} > \delta_0$. If H_0 is true, $\frac{\underline{a}'\bar{\underline{X}} - \delta_0}{\sqrt{\underline{a}'S\underline{a}}}\sqrt{n} \sim t(n-1)$.

So we reject if

$$\frac{\underline{a}'\bar{\underline{X}} - \delta_0}{\sqrt{\underline{a}'S\underline{a}}}\sqrt{n} > t_\alpha(n-1).$$

2) If H_1 is $\underline{a}'\underline{\mu} \neq \delta_0$, we reject at level $\alpha \in (0,1)$ if:

$$\frac{|\underline{a}'\bar{\underline{X}} - \delta_0|}{\sqrt{\underline{a}'S\underline{a}}}\sqrt{n} > t_{\frac{\alpha}{2}}(n-1).$$

We are not learning these formulas by heart, we are obtaining them from pivotal distributions!

Question: did we prove that:

$$(*)\quad \Pr\left[\underline{a}'\underline{\mu} \in \left[\underline{a}'\bar{\underline{X}} \pm t_{\frac{\alpha}{2}}(n-1)\sqrt{\frac{\underline{a}'S\underline{a}}{n}}\right], \forall \underline{a} \in \mathbb{R}^p\right] = 1-\alpha \quad ?$$

No, this is totally wrong. That's why we are computing "one CI at-the-time".

What is wrong is the term $t_{\frac{\alpha}{2}}(n-1)$: we need to construct a different pivotal quantity for this new problem.

II) Simultaneous CI for linear combinations for μ:

We want something like this:

$$\Pr\left[\underline{a}'\underline{\mu} \in \left[\underline{a}'\bar{\underline{X}} + (?)\sqrt{\frac{\underline{a}'S\underline{a}}{n}}\right], \forall \underline{a} \in \mathbb{R}^p\right] = 1-\alpha.$$

We are interested in simultaneous inference(*), that doesn't have to be mixed with an-at-the-time interval.

We need a small excursus on linear algebra.

Example: coin toss.

$$X_1 = \begin{cases} 0, & 1/2 \\ 1, & 1/2 \end{cases}, \quad X_2 = \begin{cases} 0, & 1/2 \\ 1, & 1/2 \end{cases} \qquad X_1, \dots, X_n \text{ iid } \text{Bern}\left(\tfrac{1}{2}\right).$$

1) $\Pr[X_i = 1] = \frac{1}{2} \quad \forall i = 1, \dots, 100.$ one-at-the-time

2) $\Pr[X_i = 1, \forall i = 1, \dots, 100] \neq \frac{1}{2}.$ simultaneously

$\hookrightarrow = \left(\frac{1}{2}\right)^{100}.$

A little excursus in Algebra:

Let $\underline{b}, \underline{d} \in \mathbb{R}^p$. $\quad \frac{\langle \underline{b}, \underline{d} \rangle}{\|\underline{b}\| \|\underline{d}\|} = \cos(\vartheta).$

$$\Rightarrow \frac{(\langle \underline{b}, \underline{d} \rangle)^2}{\|\underline{b}\|^2 \|\underline{d}\|^2} \leq 1.$$ Equality holds if $\underline{b} \in \text{span}(\underline{d})$, i.e. $\underline{b} \propto \underline{d}$.

Hence, $(\underline{b}'\underline{d})^2 \leq \left(\sum_i b_i^2\right)\left(\sum_i d_i^2\right)$ and equality holds if $\underline{b} \in \text{span}(\underline{d})$.

Writing it in a different way,

$$\int fg \leq \left(\int f^2\right)\left(\int g^2\right).$$ Cauchy-Schwartz inequality

[all of the statistics in $\mathbb{R}^p$ can be transfert in a Hilbert space]

Take now B $p \times p$ positive definite matrix.

We have that:

$$(\underline{b}'\underline{d})^2 \leq (\underline{b}' B \underline{b})(\underline{d}' B^{-1} \underline{d})$$

extended Cauchy-Schwartz (CS)

Proof:

$$\underline{b}'\underline{d} = \underline{b}' B^{1/2} B^{-\frac{1}{2}} \underline{d}.$$

$$\Rightarrow (\underline{b}'\underline{d})^2 = (\underline{b}' B^{1/2} B^{-\frac{1}{2}} \underline{d})^2 \overset{C.S.}{\leq} (\underline{b}' B^{1/2} B^{+\frac{1}{2}} \underline{b})(\underline{d}' B^{-\frac{1}{2}} B^{-\frac{1}{2}} \underline{d}')$$

$$= (\underline{b}' B \underline{b})(\underline{d}' B \underline{d}).$$

Equality holds if $\underline{b}' B^{1/2} \in \text{span}(B^{-1/2}\underline{d})$, or $\underline{b} \in \text{span}(B^{-1}\underline{d})$.

(end)

("maximum lemma")

Proposition: Let B $p \times p$ pos. def., $\underline{d} \in \mathbb{R}^p$. Then:

$$\max_{\underline{x} \in \mathbb{R}^p, \underline{x} \neq \underline{0}} \frac{(\underline{x}'\underline{d})^2}{\underline{x}' B \underline{x}} = \underline{d}' B^{-1} \underline{d}$$

maximum lemma

Proof: $(\underline{x}'\underline{d})^2 \leq (\underline{x}' B \underline{x})(\underline{d}' B^{-1} \underline{d})$.

If $\underline{x} \neq \underline{0} \Rightarrow \underline{x}' B \underline{x} > 0$.

$$\Rightarrow \frac{(\underline{x}'\underline{d})^2}{\underline{x}' B \underline{x}} \leq \underline{d}' B^{-1} \underline{d} \quad \forall \underline{x} \neq 0.$$

Equality holds if $\underline{x} \in \text{span}(B^{-1}\underline{d})$.

(end)

Back to statistics:

$\underline{X}_1, \dots, \underline{X}_n$ iid $\sim N_p(\underline{\mu}, \Sigma)$, $\bar{\underline{X}} = \frac{1}{n}\sum_{i=1}^n \underline{X}_i$.

Let $\underline{a} \in \mathbb{R}^p$, take $\frac{(\underline{a}'(\underline{X}-\underline{\mu}))^2}{\underline{a}' S \underline{a}} n$. By maximum lemma,

(Max. Lemma) (Hotelling)

$$\max_{\underline{a} \in \mathbb{R}^p} n \frac{(\underline{a}'(\underline{X}-\underline{\mu}))^2}{\underline{a}' S \underline{a}} = \underbrace{n(\underline{X}-\underline{\mu})' S^{-1} (\underline{X}-\underline{\mu})}_{\| \ T} \sim \frac{(n-1)p}{n-p} F(p, n-p).$$

$\Rightarrow$ We have our pivotal statistics! Notice that it's the same elliptical statistics that we use to build the Confidence Region. Therefore:

$$\Pr\left[\frac{|\underline{a}'(\bar{\underline{X}}-\underline{\mu})|}{\sqrt{\underline{a}' S \underline{a}}} \sqrt{n} \leq c, \ \forall \underline{a} \in \mathbb{R}^p, \underline{a} \neq \underline{0} \right] = 1-\alpha.$$

? $\longrightarrow$ key equality (★)

$$\Pr\left[\max_{\substack{\underline{a} \in \mathbb{R}^p \\ \underline{a} \neq \underline{0}}} \frac{|\underline{a}'(\bar{\underline{X}}-\underline{\mu})|}{\sqrt{\underline{a}' S \underline{a}}} \sqrt{n} \leq c \right] = 1-\alpha.$$

(★) if the quantity is smaller $\forall \underline{a}$, then it will be smaller also for the $\underline{a} \in \mathbb{R}$ generating the biggest quantity.

Taking the square (monotone function):

$$\Pr\left[\max_{\underline{a} \in \mathbb{R}^p, \underline{a} \neq \underline{0}} \frac{(\underline{a}'(\bar{\underline{X}} - \underline{\mu}))^2}{\underline{a}' \underline{\Sigma} \underline{a}} n \leq c^2 \right] = 1 - \alpha$$

$$\sim \frac{(n-1)p}{n-p} F(p, n-p) \quad \Rightarrow \quad c^2 = \frac{(n-1)p}{n-p} F_\alpha(p, n-p).$$

We can conclude that:

$$\Pr\left[\frac{|\underline{a}'(\bar{\underline{X}} - \underline{\mu})|}{\sqrt{\underline{a}' \underline{\Sigma} \underline{a}}} \sqrt{n} \leq \sqrt{\frac{(n-1)p}{n-p} F_\alpha(p, n-p)} \quad \forall \underline{a} \in \mathbb{R}^p \right] = 1 - \alpha.$$

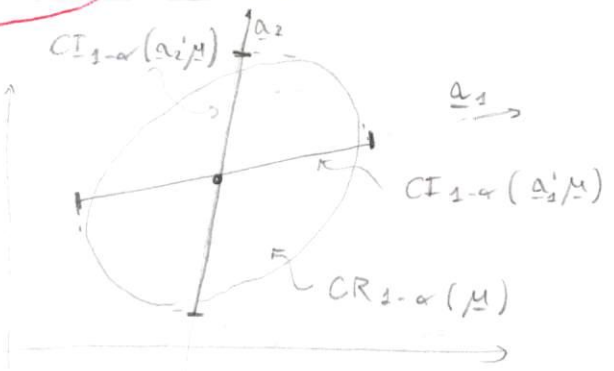
The quantile is a Fisher, not a Student-t!

$$CI_{1-\alpha}(\underline{a}'\underline{\mu}) = \left[\underline{a}' \bar{\underline{X}} \pm \sqrt{\frac{(n-1)p}{n-p} F_\alpha(p, n-p)} \sqrt{\frac{\underline{a}' \underline{\Sigma} \underline{a}}{n}} \right].$$

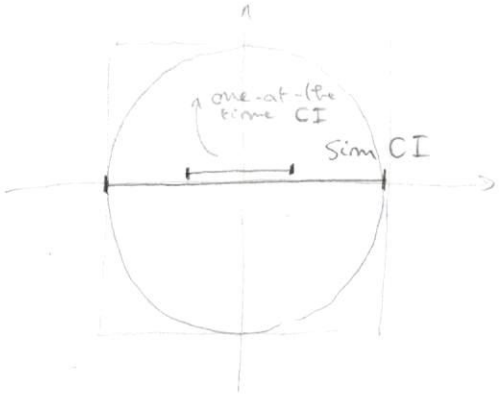
simultaneous CI for $\underline{a}'\underline{\mu}$

Geometry of the CI:

The simultaneous CI is the projections of the Confidence Region of $\underline{\mu}$ along all the directions $\underline{a}$.



Comment: today, with Big Data, we want to estimate a large number of parameters. Simultaneous CI is hence very used.



We have two "extremes":

- simultaneous cover all possible linear combinations: larger CIs.
- one-at-the-time CIs: too specific.

What if we specify just a finite number of them? Your client wants to see a smaller CI, so we are tempted to sell them the one-at-the-time CI. The intermediate method is given by Bonferroni.